Supporting Highly Mobile Users in Cost-Effective Decentralized Mobile Operator Networks

Tarik Taleb, *Senior Member, IEEE*, Konstantinos Samdanis, *Member, IEEE*, and Adlen Ksentini, *Member, IEEE*

***Abstract*—Due to the tremendous increase in mobile data traffic, there is a general trend toward the decentralization of mobile operator networks, at least to a certain extent. This shall be further facilitated with the virtualization of mobile network functions and the enabling of mobile cloud networking, whereby mobile networks are created on demand and in a flexible manner. Mobile network decentralization will not be efficient without rethinking mobility management schemes, particularly for users moving for a long distance and/or at a high speed (e.g., vehicles and bullet trains). To support such highly mobile users, this paper introduces: 1) a data anchor gateway (GW) relocation method based on user mobility, history information, and user activity patterns, and 2) a handover management policy that selects a target base station or evolved Node B (eNB) in a way to minimize mobility anchor GW relocation. The performance of the proposed schemes is evaluated via Markov model-based analysis and through simulations. Encouraging results are obtained, validating the design objectives of the scheme.**

***Index Terms*—Cellular networks, markov chains, mobile radio mobility management, QoS, 4G mobile connection.**

I. Introduction

As MOBILE networking and services are entering a new communication era offering smartphones to users with higher capabilities and more diverse applications, emerging service requirements create new challenges for the current mobile network architecture. Such new requirements partly reflect the popularity of several new services and the emerging content-rich and bandwidth-intensive mobile applications. In addition, they capture the operator's desire to offer flat-rate tarifs to attract more users, encouraging the adoption of new services. Such a business paradigm may work great in an early phase assisting the success of new technologies, i.e., Long-Term Evolution (LTE), but at a later stage, it may create a rebound effect with serious revenue problems for operators. Indeed mobile operators are facing a challenging task of how to accommodate huge traffic volumes, far beyond the original network capacity [1], [2].

Manuscript received July 15, 2013; revised November 22, 2013; accepted December 30, 2013. Date of publication January 10, 2014; date of current version September 11, 2014. The review of this paper was coordinated by Prof. Y.-B. Lin.

T. Taleb and K. Samdanis are with NEC Europe, Ltd., Heidelberg 69115, Germany (e-mail: tarik.taleb@neclab.eu; samdanis@neclab.eu).

A. Ksentini is with the Institut de Recherche en Informatique et Systèmes Aléatoires, University of Rennes 1, Rennes 35042, France (e-mail: adlen.ksentini@irisa.fr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TVT.2014.2299633

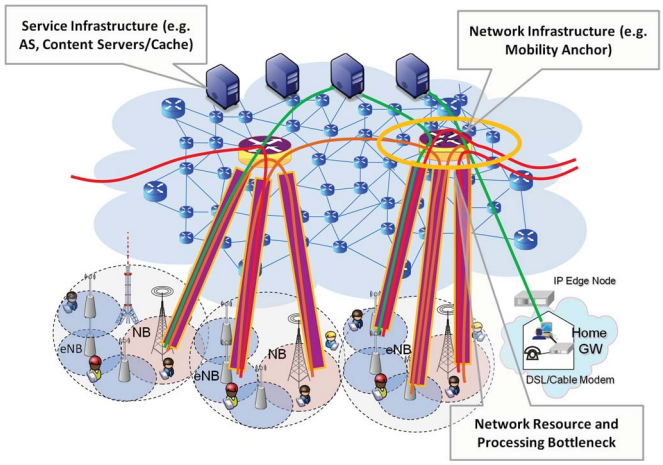

Fig. 1. Today's centralized mobile operator networks.

Effectively, such new requirements challenge the current mobile network architecture, which is highly centralized and not optimized for high-volume data applications. The main problem relates to the fact that central gateways (GWs) handle all mobile traffic, acting as a data and mobility anchor for several radio access points without any complementary caching or data offload support at the network edge. Fig. 1 shows the current centralized network architecture pointing out its shortcomings, which include the following:

- high concentration of traffic demands toward central network locations consuming high backhaul resources in terms of bandwidth and imposing higher processing demands on centralized mobility GWs, easily leading to undesirable bottlenecks;
- long communication paths between users and servers leading to waste in core network and backhaul resources, undesirable delay, and poor quality of experience (QoE);
- higher risk in network availability since a centralized architecture creates single locations of failure.

A straightforward solution for mobile operators is to invest in upgrading their network infrastructure in terms of backhaul speed and core network resources with the objective of always being able to accommodate peak-hour traffic demands. While these are technical-wise feasible solutions, financially, they are challenging, particularly due to the modest average revenues per user, given in turn the trend toward flat-rate business models. Operators are thus interested in cost-effective methods for accommodating the ever-increasing mobile network traffic ensuring minimal investment into the current infrastructure.

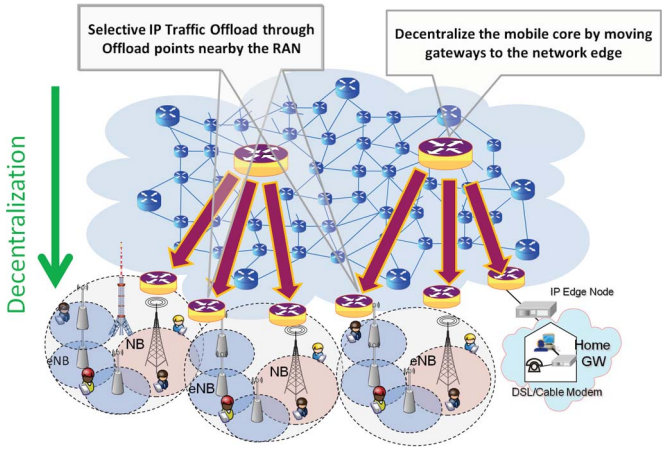

Fig. 2. Trend toward decentralized mobile operator's networks.

Network decentralization is a key enabler that allows operators to be equipped with economically competitive solutions against increased traffic demands and flat-rate charges. The basis for realizing network decentralization is to place small-scale network nodes with mobility and IP access functionalities, similar to those provided by the currently centralized GWs, toward the network edge. Such local data anchor GWs allow operators to employ solutions that can selectively offload traffic as close to the radio access network (RAN) as possible [3]. Such vision is also in line with the quest for a network architecture flatter than what has been achieved with the Evolved Packet Core (EPC) [4], [5]. In principle, decentralized networks may support the following network entities being locally deployed, including packet data network GWs (PDN-GWs or P-GWs), serving GWs (S-GWs), and mobility management entities (MMEs), with the objective of serving a local community of users. An example that demonstrates the vision behind the decentralized mobile network deployment is shown in Fig. 2. Effectively, by breaking out selected traffic at entities close to the moving terminal, operators will be able to avoid overloading their scarce core network resources.

The discussions and analysis in Third-Generation Partnership Project (3GPP) have been on the definition of the architecture, i.e., on where the point of local breakout/traffic offload should be placed, in addition to issues regarding security, charging, mobility, traffic control/handling [6], and optimal GW selection [7]. The efficient usage of the network resources in such a decentralized architecture requires the selection of optimal data anchor GWs for user equipment (UE) [7]. The notion of GW optimality largely depends on operators' policies. It can be in terms of geographical proximity of a user to a GW and/or GW load [7], [14], [27], [28], [30]. It can also depend on the service/application type used by the users (e.g., a GW can be optimal for anchoring Facebook traffic but not for YouTube traffic) [30]. In current standards, notifying UE, particularly in idle mode, of the availability of such an optimal GW is immediately followed by enforcing the UE to disconnect and reconnect to the network; during this operation, the UE is relocated to the currently optimal GW. For UE in idle mode, moving fast, and/or for a long distance (e.g., vehicles) [31], this solution may lead to frequent unnecessary GW relocations. To resolve this issue, a number of solutions can be envisioned. The objective of this paper is to compare among the different solutions and discuss their advantages and pitfalls.

Another concern in decentralized mobile networks, particularly for highly mobile users, such as those on board bullet trains, consists in the fact that serving areas and pool areas of S-GWs and MMEs, respectively, get smaller and that is for the purpose of localizing mobility management [8]. This, in combination with the fact that UE tends to be always active due to many (e.g., cloud) applications running in the background, as in smartphones or vehicles with LTE access, increases the likelihood of performing handoffs with S-GW and/or MME relocation. Handovers with S-GW/MME relocation may impact the QoS of an ongoing session since they incur additional delay compared with the non-S-GW/MME relocation handover counterparts. In addition, MME/S-GW relocation requires the establishment of sessions/bearers along the new path, incurring additional overhead in terms of signaling. Issues pertaining to admission control at nodes along the new path may also occur. Given these reasons and more, the 3GPP specifications [4] indicate that S-GW/MME relocation during handoff for UE in evolved packet system (EPS) connection management connected mode (i.e., active mode) is to be avoided when possible. However, due to lack of coordination between the RAN and core network nodes, the possibility of S-GW/MME relocation is not taken into account in the handover decision at the RAN.

Optimizing the handover decision by combining intelligence from both the RAN and the core network to avoid S-GW/MME relocation, when possible, is one of the objectives of this paper. Such handover optimization can improve handover delay and reduce associated signaling, positively affecting both the network performance and the user's perceived quality. The importance of the proposed handoff optimization is significant in case of a high number of active UE with high mobility features, such as LTE-connected vehicles or smartphones on board moving objects, particularly when the S-GW service areas/MME pool areas are of small size (due to the foreseen mobile network decentralization as previously mentioned). Its importance is also significant when handling the mobility of a large group of UE (e.g., all performing simultaneous handoffs with S-GW/MME relocation due to a specific event such as concert, arrival of train at a station, etc). While one may argue that the example of train arrivals at a station can be taken into account during the mobile network planning phase, there are scenarios whereby the initial network planning becomes no longer optimal (such as due to the construction of a new road or a shopping center) and frequent handovers with S-GW/MME relocation may then become inevitable.

The remainder of this paper is structured as follows. Section II describes some related work. Section III introduces our proposed solutions to the two aforementioned issues. Section IV develops an analytical model for the proposed solutions. The performance evaluation of the proposed solutions, which is based on both the analytical model and simulations, is introduced and discussed in Section V. Section VI concludes this paper with a summary recapping the main findings of this paper.

TABLE I
LTE NETWORK ENTITIES AND THEIR FUNCTIONS

Node	Description
eNodeB	Evolved NodeB, the LTE base station.
MME	Mobility Management Entity, a control plane entity for all mobility related functions, paging, authentication, bearer management in the Evolved Packet System
P-GW	Packet Data Network Gateway, interfaces with the Packet Data Network (e.g., Internet)
S-GW	Serving Gateway, Local mobility anchor for intra-3GPP handoffs.
HSS	Home Subscriber Server, main database containing subscription-related information.

II. RELATED WORK

Before delving into the review and analysis of some related work, we first introduce the functionality of the most important network entities used by the LTE reference architecture, namely EPS, as listed in Table I.

With the introduction of the 3GPP LTE, certain network decentralization features focusing both on network architecture and network management were brought forward, significantly improving the prior universal mobile telecommunications system (UMTS). Specifically, considering the network architecture, LTE merged the former radio network controller (RNC) within Node B, introducing a new radio access element called evolved Node B (eNB) that flattens and simplifies the prior UMTS architecture [4]. In terms of network management, LTE advances the prior UMTS-based configuration and optimization methods toward a distributed self-organized paradigm [9]. A further step toward an even flatter UMTS architecture is in [10], whereby the GW GPRS support node (GGSN) and serving GPRS support node (SGSN) are integrated in NodeBs, whereas in [11], a similar approach is introduced in the context of EPS with the aim of bringing GW functions to the edge of the network, merging in this way the radio access and core networks.

In principle, there is a fundamental technology- and cost-related tradeoff behind the adoption of either centralized or distributed network architecture. Costly network equipment are usually shared, creating centralized architectures. This was the case in the initial phase of third-generation (3G) deployment, wherein a centralized architecture was preferred to share core network utilities and processing resources, while keeping the cost of base stations low [10]. Nowadays, the evolution of computer technology has significantly reduced the cost of equipment, advancing their deployment flexibility. However, the ever-increasing mobile data volumes render the network utilization cost high, creating significant revenue problems for operators. For this reason, operators have been looking into data offloading solutions, introducing P/S-GW functionality toward the network edge, i.e., close to eNBs. Additionally, the provisioning of content-distribution-network services via caches placed nearby the network edge may complement data offloading, reducing further network cost as within the backhaul and via a more efficient resource usage [12].

A tutorial regarding the data offloading techniques, including local IP access (LIPA), selective IP traffic offload (SIPTO), and IP flow mobility and seamless offload, focusing on 3GPP Rel-10 is available in [13], whereas further details on LIPA/SIPTO data offloading are provided in [14], which shows specific network architectures and service requirements that meet the current decentralized needs, enlightening also network management and deployment issues with emphasis on QoS and service continuity. A complementary analysis on the architecture and main benefits associated with the use of decentralized architectures from the Internet Engineering Task Force (IETF) perspective is documented in [15] and [16], with the most important features being route optimization and increased robustness.

Current research efforts regarding decentralized cellular networks focus on mobility and service continuity, and on efficient resource management related to GW selection and relocation. Ensuring service continuity and QoE for active users while reducing signaling, by avoiding the use of core network equipment, is the ultimate goal of decentralized mobile networks. Nevertheless, the means of achieving decentralization is distributed mobility management. A study that examines and compares centralized and distributed mobility in the context of 3GPP is presented in [17], also elaborating on a dynamic and distributed mobility management solution, which merges the mobility anchor and base station. In particular, the support of distributed mobility is realized with the use of tunneling to forward traffic upon a handover, also allowing users to establish flows via different mobility anchors for efficient resource usage. This latter session establishment mechanism comprises a significant resource optimization feature, which is also followed in [7], with complementary mechanisms for efficient GW selection considering load balancing, service availability, and QoS. The fundamental concepts of GW-based load balancing with respect to user performance are analyzed in [18], where an inter-GW load balancing protocol is presented and evaluated considering a GW pool in association with a group of base stations. A more advanced solution for providing service continuity is introduced in [19]. Such a solution distributes the content of a centralized mobility anchor over a set of distributed mobility agents utilizing the concept of virtual routers and distributed hash tables, achieving substantial benefits in load balancing and resiliency.

A comprehensive study on location management solutions for legacy mobile networks is provided in [20], considering different mobility models. In this paper, we provide a similar study envisioning random mobility for decentralized 3GPP LTE networks. Specifically, so far, none of the state-of-the-art network decentralization solutions considered the case of P-GW and S-GW relocation within the 3GPP LTE for idle and active users. In contrast, as will be explained in the following, our approaches, proposed herein, consider the selection of optimal P-GWs for high-speed users in a decentralized mobile network environment. This is a fundamental difference because the prior approach updates the location area to enable the network to locate the user when traffic is routed toward him or her, whereas our proposed approaches update the data anchor point, namely P-GW, to enable high-speed users, depending on their

activity level, to obtain optimal connectivity from both the user and network perspectives. The contributions of this paper are therefore twofold.

- The first contribution relates to the notification of optimal P-GW availability and enforcement of P/S-GW relocation for highly mobile UE in idle mode.
- The second contribution relates to a solution for S-GW/MME relocation avoidance via optimal target eNB selection during a handover decision for highly mobile UE in active mode.

Considering the PDN connectivity of idle users, the work in [7] introduces the main challenges associated with maintaining PDN connectivity. It also proposes and analyzes some of the PDN connection reestablishment solutions proposed in this paper, namely, solutions for periodic PDN connection reestablishment, PDN connection reestablishment upon a tracking area update, and PDN connection reestablishment based on network indication. This paper complements the prior discussed solutions introducing further parameters to perform PDN connection reestablishment based on user mobility and history information, and on user activity patterns. User mobility prediction is also considered in [22] focusing on ensuring QoS of multimedia applications by minimizing the frequency and magnitude of fluctuations. The method introduces a stochastic path prediction model based on history information assuming *a priori* knowledge of the destination. Such a history model could potentially be adopted in this paper to assist with mobility and user activity prediction.

Although such selection criteria have been encountered before for centralized schemes, as in [23], considering mobile IP networks, their scope was mostly focused on active sessions. Their performance is expected to be different in decentralized arrangements centered on idle users. The reason is mainly the different focus that concentrates on the timing issues of the new session establishment and whether the "always on" connectivity of UE is maintained, i.e., parameters that affect the signaling overhead and session establishment delay. This paper also introduces an analytical framework using Markov-based models to provide a performance comparison of the proposed solutions.

For highly mobile users, i.e., LTE-connected vehicles on a highway and users on board bullet trains, network decentralization would also have a significant impact on the standardized tracking area update (TAU) procedure and on the maintenance of idle-mode PDN connections. Although the TAU approach as in [8] encounters a degree of TAU overlapping, allocating to UE a GW in the center of the TAU area that could potentially serve them for a wide geographical area considering the position of the initially associated eNB, this approach mostly holds for centralized architectures. In decentralized schemes, we envision smaller service areas and MME pool areas; thus, additional mechanisms to reduce the frequency of service area/MME pool area relocations are recommended. In [21], a self-organized method, which adopts the tracking areas according to the user mobility taking into account long-term history data, is introduced based on graph partitioning heuristics. In this paper, we complement such methods by introducing a handover management policy that selects neighboring eNBs to avoid service area relocations whenever possible. The proposed handover policy is also equipped with a hysteresis mechanism to avoid frequent handovers related to UE located at the edge of a cell and exposed to ping-pong effects.

III. Proposed Solutions for the Support of Highly Mobile Users in Decentralized Mobile Networks

Before elaborating on the details of our solutions toward supporting highly mobile nodes in decentralized networks, we stress that, while the description in this paper relates to evolved UMTS terrestrial radio access network (eUTRAN), the same applies to UMTS and other types of mobile networks. In the case of UMTS, MMEs, eNBs, S-GWs, P-GWs, and service areas map onto SGSNs, base stations or RNCs, SGSNs, GGSNs, and routing areas or location areas, respectively.

A. Smart P/S-GW Relocation for Highly Mobile UE in Idle Mode

As stated earlier, the current 3GPP architecture is centralized with relatively few P-GWs, which serve a high population of UE within a broad geographical area. In such an arrangement, the network, specifically MMEs, may disconnect UE whenever an optimal P-GW becomes available without considering any user-related intelligent information but simply following certain static predetermined rules, e.g., based on geographical distance. For specific UE in idle mode, P-GW relocation may occur by having the network disconnect the UE and immediately triggers the UE to reconnect to the network. Some applications, running on the UE, may then be forced to reregister/subscribe again [e.g., instant messaging or applications based on the session initiation protocol (SIP)]. In such case, the state of applications when they refresh their registration/subscription and the impact of the forced disconnection may potentially cause disruptions on ongoing communications. In case P-GW relocation is triggered for every handover conducted by the UE, the user will be always connected to the optimal P-GW, but this comes at significant signaling overhead. Of particular interest, for fast/far-moving UE (e.g., users using smartphones on board bullet trains), this will clearly cause a large number of unnecessary signaling that shall waste network resources, not to mention the negative impact on the UE battery lifetime. Admittedly, such a per-handover P-GW relocation solution, referred to as baseline GW relocation (or Solution 1) throughout this paper, is straightforward and has minimal impact on the MME. It requires no knowledge (at the MME) about the user behavior nor about any information regarding the type of applications running on the UE.

Fig. 3 shows schematic flowcharts of four alternative solutions to the aforementioned baseline GW relocation method. As the first alternative solution (to Solution 1), the MME may disconnect the UE with some intelligent logic. In particular, in one solution, the MME could take into account the history of UE mobility (e.g., number of handoffs/TAUs performed during a specific period) to decide whether to immediately disconnect

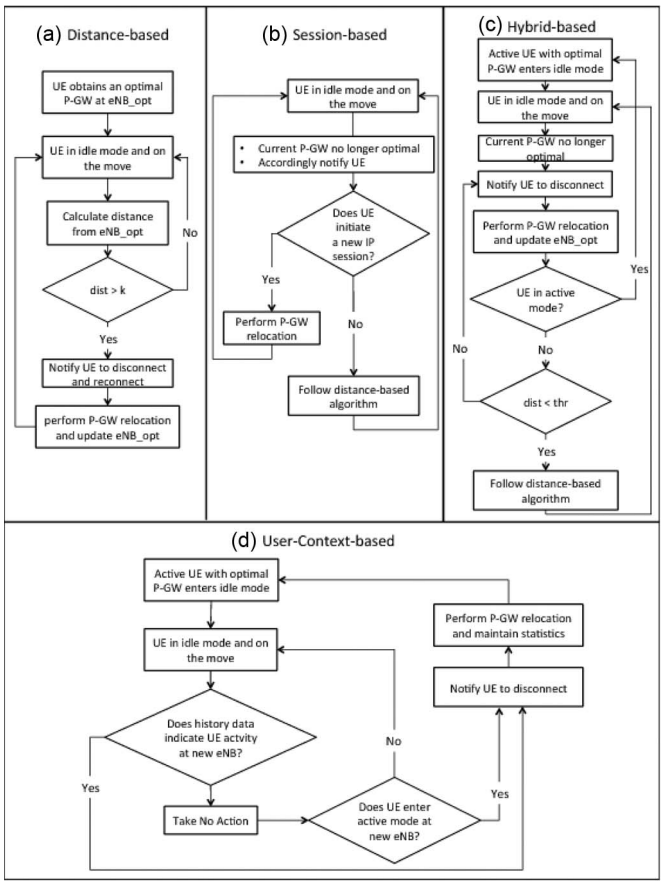

Fig. 3. Schematic flowcharts of the proposed solutions. (a) Distance based. (b) Session based. (c) Hybrid based. (d) User context based.

the UE or to delay the disconnection request until after a predetermined period and/or after the occurrence of a number of handoffs and/or TAUs (e.g., when the UE migrates to a cell k hops far away from the initial cell, which the UE was associated at the time of the PDN connection setup). In comparison to the described baseline GW relocation approach (i.e., Solution 1), this UE mobility-aware GW relocation method [referred to as Solution 2 throughout this paper; see Fig. 3(a)] reduces the frequency of PDN disconnection requests and consequently has less impact on the UE battery lifetime. However, the solution intuitively relies on some intelligent logic at the MME as the MME needs to keep track of UE's mobility.

In a further alternative solution, the MME may indicate the possible availability of an optimal P-GW in the TAU response without enforcing the disconnection. The disconnection is enforced only at specific areas, such as service area boundaries. Compared with the baseline GW relocation and the UE mobility-aware GW relocation approaches, this solution [referred to as Solution 3 throughout this paper; see Fig. 3(b)] is clearly efficient in terms of reducing signaling as it enforces PDN disconnection only at service area boundaries and does not require any intelligent logic, such as considering the UE mobility pattern at the MME (i.e., simple implementation). Furthermore, it provides UE with the flexibility to decide when to request the establishment of a new PDN connection (via the optimal P-GW) based on the type of applications being active and their state. Indeed, this flexibility allows the UE,

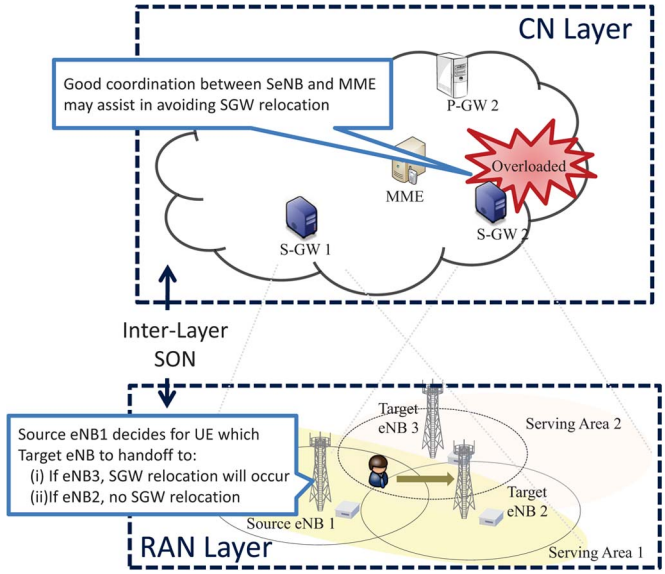

Fig. 4. Lack of coordination among the core–RAN and S-GW relocations.

for example, to reestablish the PDN connection—if it was previously indicated by the network—when the application updates its periodic registration/subscription status, and not just after, which would require another reregistration/subscription. In another approach combining Solutions 2 and 3 [and referred to as Solution 4 throughout this paper; see Fig. 3(c)], the MME may disconnect UE for a given number of times as in the baseline GW relocation method (i.e., Solution 1), and after that, the MME indicates the possible availability of an optimal P-GW in the TAU response without enforcing the disconnection (as in Solution 3). In Solution 5, employing some level of context-awareness [see Fig. 3(d)], UE devices with a history record for frequently initializing IP sessions during certain periods and/or at certain locations are always connected to the optimal P-GW during these periods of time and/or when the UE devices are at these specific locations. Other UE devices are disconnected and reconnected to an optimal P-GW only when they initiate an IP session.

B. S-GW/MME Relocation Avoidance During Handover for UE in Active Mode

As stated earlier, from the core network point of view, it is highly preferable to avoid MME/S-GW relocation for UE in active mode [4]. However, the handover decision is a pure RAN decision. Fig. 4 shows the issue, in which UE, being currently connected to eNB1, is about to perform handoff. Given the current location of the UE, the UE has two possible target eNBs, namely eNB2 and eNB3, as shown in Fig. 4. The two eNBs are in two different service areas: eNB2 is in the same service area 1, serviced by S-GW1, as source eNB1, and eNB3 is in a different service area 2, serviced by S-GW2. After receiving measurement reports from the UE, source eNB1 decides for the UE which target eNB to hand over to. In case eNB2 is selected, the S-GW relocation can be avoided. However, if eNB3 is selected, S-GW relocation from S-GW1 to S-GW2 will be inevitable. The issue becomes further significant in the case that S-GW2 is overloaded and shall be avoided. It should

be noted that the addressed problem is just an example of the conflict that may happen between functions at the RAN and the core network, when they are run individually. A handover optimization function that reflects S-GW relocation avoidance in the handover decision is thus of vital importance, not only for the UE's QoE but also for supporting the GW selection functions of the core network. It shall be noted that the S-GW relocation issue may be also experienced in the case of handoff with MME relocation (i.e., S-GW and S-GW service areas can be replaced by MME and MME pool areas, respectively) [8]. Admittedly, due to the relatively larger size of MME pool areas (in comparison to S-GW service areas), handoffs with MME relocation may be less frequent than handoffs with S-GW relocation. We thus focus in this paper on handoffs with S-GW relocation.

In general, a UE keeps sending measurement reports to the eNB that it is currently connecting to (i.e., source eNB). These measurement reports contain information about the detected signal strengths from the different neighboring eNBs. Taking into account the reported signaling strengths, in addition to information regarding the load of each neighboring eNB, the source eNB decides which eNB the UE shall hand off to (i.e., target eNB), encountering also the aforementioned issue. In general, we propose that RAN, particularly the source eNB, takes into account the possibility of S-GW/MME relocation in the handoff decision based on a good coordination between the RAN nodes and the core network nodes. This could be achieved by having the information on the possibility of S-GW/MME relocation available *a priori* to source eNB. Based on that, the source eNB decides on optimal target eNB to avoid MME or S-GW relocation. In the case of static S-GW service areas (e.g., S-GW is not UE specific), eNBs on the edge of each service area can be preconfigured by operation and management with such information. Alternatively, such information can be assessed by source eNB from the "distance" to the current S-GW or can be explicitly indicated *a priori* by MME to the source eNB when the UE handoffs to the source eNB. This indication from MME is done only when required.

In an alternative solution, source eNB provides the MME, e.g., in decreasing order of radio quality, a list of possible target eNBs. The MME then recommends the best eNB to avoid S-GW/MME relocation. It shall be noted that the source eNB provides the MME with such feedback in case all possible target eNBs have the same characteristics in terms of signal strength, load, and/or other contextual information. Consulting the MME could be also performed only when required by the MME, i.e., based on explicit indication/triggering.

IV. Analytical Model

A. System Models

Here, we model the solution variants given earlier, including Solutions 1, 2, 3, and 4, but not Solution 5, which is considered later in the simulation study. The aim of these models is to quantify the frequency of users in being connected to their relevant optimal P-GWs and the cost associated with disconnecting UE devices once they are relocated from the prior optimal P-GWs. The system is analyzed using Markovian models. We assume

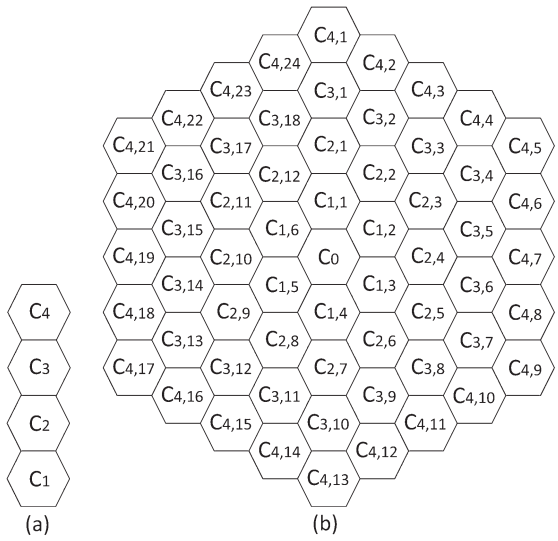

Fig. 5. Envisioned mobility models. (a) 1-D (linear) model. (b) 2-D model.

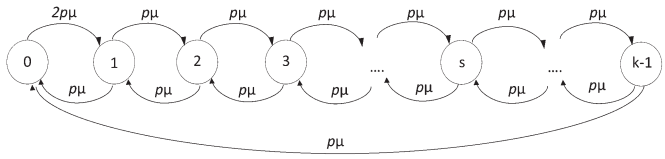

Fig. 6. Modeling Solution 2 using the 1-D mobility model.

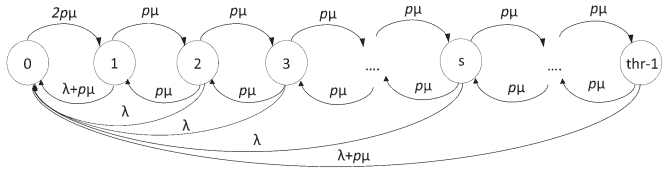

Fig. 7. Modeling Solution 3 using the 1-D mobility model.

that the 3GPP network is divided into hexagonal cells and that each cell represents an eNB wherein a P-GW is collocated. Hence, each cell can be associated with a different P-GW. Two random mobility models were used, namely the 1-D and 2-D models. It shall be stated that since the distance between two neighboring cells is usually on the order of few kilometers, such random mobility model may be easily applicable to high-speed vehicles. While the 1-D model is used when the UE mobility is limited to a prespecified unidirectional trajectory, such as roads, trains, and tunnels, the 2-D model suits better urban areas where UE can move in any direction without restriction. Fig. 5 shows both mobility models. The obtained results will be then employed to express the performance metrics in terms of the probability to be connected to the optimal P-GW and the cost of P-GW relocation.

1) 1-D Mobility Model: In this model, UE moves from one cell to another cell with the same probability $p = 1/2$ (i.e., as there are only two possible destinations). Let $X(t)$ denote the distance at instant time t from the UE's location to the optimal P-GW in terms of number of hops. The residence time of the UE in each cell follows an exponential distribution with a mean $1/\mu$. Furthermore, the arrival rate of an application session in the cell is assumed to follow a Poisson distribution with a mean

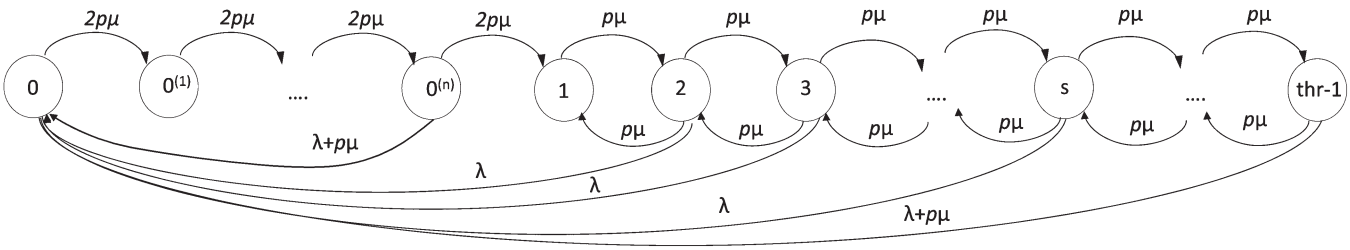

Fig. 8. Modeling Solution 4 using the 1-D mobility model.

λ. Thus, the interarrival time of an application session into a cell follows an exponential distribution with a mean $1/\lambda$. In the following, we use this 1-D mobility model to model the given solutions.

In Solution 2 (distance based), UE is disconnected once it performs a handover toward an eNB with a distance k hops or more away from the one associated with the optimal P-GW. In this case, the system $\{X(t), t \geq 0\}$ forms a continuous-time Markov chain (CMTC) with the state space $\{S_1 = 0, 1, 2, \ldots, (k-1)\}$, as shown in Fig. 6. Let $\pi_s = \lim_{t\to\infty} \Pr[X(t) = s]$, $s \in S_1$, be the stationary probability distribution of $X(t)$. The balance equations to derive the stationary probability are given as follows:

$$\begin{cases} 2p\mu\pi_0 = p\mu\pi_1 + p\mu\pi_{k-1} \\ 2p\mu\pi_1 = p\mu\pi_0 + p\mu\pi_2 \\ 2p\mu\pi_s = p\mu\pi_{s-1} + p\mu\pi_{s+1} \\ 2p\mu\pi_{k-1} = p\mu\pi_{k-2} \\ \sum_{s=0}^{k-1} \pi_s = 1. \end{cases} \tag{1}$$

In Solution 3 (session based), UE is disconnected upon initiating a connection while not having an optimal P-GW. Similar to Solution 2, the system $\{X(t), t \geq 0\}$ forms a CMTC with the state space $\{S_2 = 0, 1, 2, \ldots, (\text{thr}-1)\}$, as shown in Fig. 7. The thr represents the size of the serving area (in terms of number of hops), which means that, upon reaching thr, the UE is disconnected even if it does not initiate a session. Let $\pi_s = \lim_{t\to\infty} \Pr[X(t) = s]$, $s \in S_2$, be the stationary probability distribution of $X(t)$. As shown in Fig. 7, the balance equations to derive the stationary probability are expressed as follows:

$$\begin{cases} 2p\mu\pi_0 = (\lambda + p\mu)\pi_1 + \lambda \sum_{i=2}^{\text{thr}-2} \pi_i + (\lambda + p\mu)\pi_{\text{thr}-1} \\ (\lambda + 2p\mu)\pi_1 = 2p\mu\pi_0 + p\mu\pi_2 \\ (\lambda + 2p\mu)\pi_s = p\mu\pi_{s-1} + p\mu\pi_{s+1} \\ (\lambda + 2p\mu)\pi_{\text{thr}-1} = p\mu\pi_{\text{thr}-2} \\ \sum_{s=0}^{\text{thr}-1} \pi_s = 1. \end{cases} \tag{2}$$

It shall be noted that the model of Solution 2 is a special case of Solution 3 when thr $= k$ and $\lambda = 0$.

In Solution 4 (hybrid based), UE maintains the optimal P-GW for n consecutive times. The UE is disconnected when initiating a new session while not connecting to the optimal P-GW. As in the precedent solutions, the system $X(t)$, $t \geq 0$, forms a CMTC with the state space $\{S_3 = 0, 0^{(1)}, 0^{(2)}, \ldots, 0^{(n)}, 1, 2, \ldots, (\text{thr}-1)\}$, as shown in Fig. 8. The subchain $\{0, 0^1, 0^2, \ldots, 0^{(n)}\}$ represents the case of n consecutive times that the optimal P-GW is maintained after each handoff. Let $\pi_s = \lim_{t\to\infty} \Pr[X(t) = s]$, $s \in S_3$, be the stationary probability distribution of $X(t)$. In Fig. 8, the balance equations to derive the stationary probability are given as follows:

$$\begin{cases} 2p\mu\pi_0 = (\lambda + p\mu)\pi_1 + \lambda \sum_{i=2}^{\text{thr}-2} \pi_i + (\lambda + p\mu)\pi_{\text{thr}-1} \\ \pi_{0^{(1)}} = \pi_0 \\ \pi_{0^{(2)}} = \pi_{0^{(1)}} \\ \vdots \\ \pi_{0^{(n-1)}} = \pi_{0^{(n)}} \\ (\lambda + 2p\mu)\pi_1 = 2p\mu\pi_0 + p\mu\pi_2 \\ (\lambda + 2p\mu)\pi_s = p\mu\pi_{s-1} + p\mu\pi_{s+1} \\ (\lambda + 2p\mu)\pi_{\text{thr}-1} = p\mu\pi_{\text{thr}-2} \\ \sum_{s=0}^{\text{thr}-1} \pi_s + \sum_{i=1}^{n+1} \pi_{0^i} = 1. \end{cases} \tag{3}$$

2) 2-D Mobility Model: Unlike the 1-D model, in the 2-D model, UE devices have the possibility of visiting six neighboring cells. The probability that UE moves to one of these cells is $p = 1/6$. Fig. 5(b) shows a service area where thr $= 5$. Each service area contains $(\text{thr} - 1)$ rings of cells. As in [24] and [25], each cell is represented by its ring label and its position in this ring. For instance, cells in ring k are denoted $C_{k,j} = (1 \leq j \leq 6k)$. It is worth mentioning that the ring label represents the distance of UE from its optimal P-GW.

Similar to the 1-D model, let $X(t)$ be the distance at instant time t from the UE's location to the optimal P-GW in terms of number of hops. The residence time of the UE in each cell $C_{i,j}$ follows an exponential distribution with a mean $1/\mu$. Furthermore, the arrival rate of an application session in the cell follows a Poisson distribution with a mean λ. Thus, the system $\{X(t), t \geq 0\}$ forms a CMTC with the state space $S_4 = \{C_{i,j} | 0 \leq i \leq (\text{thr}-1), 1 \leq j \leq 6i\}$. As it is defined, this chain undergoes state-space explosion problem, particularly if the thr value is high. Accordingly, as in [24] and [25], we propose reducing the state space by aggregating states that show the same behavior. We obtain a new chain, noted $A(t)$ with lower number of states. To do so, we take advantage of the symmetry of the 2-D model. In Fig. 5(b), we see that UE devices in the first ring have the same behavior and can move to each neighbor cells with the same probability. That is, UE devices come back to the cell with optimal P-GW with probability p, stay in the same ring (same distance from the optimal P-GW) with probability $2p$, and move to ring 2 (increase the distance from the optimal P-GW) with probability $3p$.

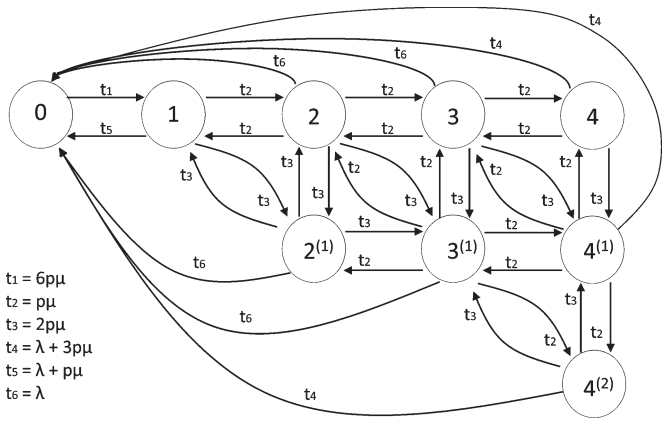

Fig. 9. Modeling Solutions 2 and 3 using the 2-D mobility model.

Thereby, all states of ring 1 can be aggregated into one state. Regarding the second ring, we differentiate between two cases. The first one is if the UE leaves the service area with probability $3p$ instead of $2p$ in the second case. Therefore, we obtain two aggregated states: State $C^*_{2,0}$ aggregates states $\{C_{2,1}, C_{2,3}, C_{2,5}, C_{2,7}, C_{2,9}, C_{2,11}\}$, and $C^*_{2,1}$ aggregates states $\{C_{2,2}, C_{2,4}, C_{2,6}, C_{2,8}, C_{2,10}, C_{2,12}\}$. Based on the algorithm presented in [24], we obtain the following aggregated states for the case of thr $= 5$, and note that the cell notation ring is organized in a clockwise manner, beginning with cell $C_{i,1}$ at the top of the ring

$$C^*_0 = C_0 = \{C_{0,0}\}$$

$$C^*_1 = C_1 = \{C_{1,1}, C_{1,2}, C_{1,3}, C_{1,4}, C_{1,5}, C_{1,6}\}$$

$$C^*_{2,0} = C_2 = \{C_{2,1}, C_{2,3}, C_{2,5}, C_{2,7}, C_{2,9}, C_{2,11}\}$$

$$C^*_{2,1} = C^{(1)}_2 = \{C_{2,1}, C_{2,3}, C_{2,5}, C_{2,7}, C_{2,9}, C_{2,11}\}$$

$$C^*_{3,0} = C_3 = \{C_{3,1}, C_{3,4}, C_{3,7}, C_{3,10}, C_{3,13}, C_{3,16}\}$$

$$C^*_{3,1} = C^{(1)}_3 = \{C_{3,2}, C_{3,3}, C_{3,5}, C_{3,6}, C_{3,8}, C_{3,9}, C_{3,11}, C_{3,12}, C_{3,14}, C_{3,15}, C_{3,17}, C_{3,18}\}$$

$$C^*_{4,0} = C_4 = \{C_{4,1}, C_{4,5}, C_{4,9}, C_{4,13}, C_{4,17}, C_{4,21}\}$$

$$C^*_{4,1} = C^{(1)}_4 = \{C_{4,2}, C_{4,4}, C_{4,6}, C_{4,8}, C_{4,10}, C_{4,12}, C_{4,14,}, C_{4,16}, C_{4,18}, C_{4,20}, C_{4,22}, C_{4,24}\}$$

$$C^*_{4,2} = C^{(2)}_4 = \{C_{4,3}, C_{4,7}, C_{4,11}, C_{4,15}, C_{4,19}, C_{4,23}\}.$$

As proven in [24], the new aggregated chain $A(t)$, raised from the initial Markovian chain $X(t)$, is also Markovian. Consequently, the system in the case of the 2-D mobility model is also forming a CMTC. In the following, we model Solutions 2, 3, and 4 using the 2-D mobility model. As stated earlier, Solution 2 is a special case of Solution 3 when $\lambda = 0$ and thr $= k$. Therefore, we consider only one model for Solutions 2 and 3. Fig. 9 shows the transition diagram of the aggregated Markov chain when thr $= 5$. Based on this figure, we can derive the stationary probability of the aggregated states C_i and C^m_i, respectively. The balance equations to solve the system are as follows:

$$\begin{cases} \pi_0 = \frac{\lambda+p\mu}{\mu}\pi_1 + \frac{\lambda}{\mu}\sum_{j=2}^{\text{thr}-2}\pi_j + \frac{\lambda}{\mu}\sum_{l=2}^{\text{thr}-2}\sum_{j=1}^{\lceil\frac{l-1}{2}\rceil}\pi_l^{(j)} \\ \quad + \frac{\lambda+3p\mu}{\mu}\pi_{\text{thr}-1} + \frac{\lambda+2p\mu}{\mu}\sum_{j=1}^{\lceil\frac{\text{thr}-2}{2}\rceil}\pi^{(j)}_{\text{thr}-1} \\ \pi_1 = \frac{6p\mu}{\lambda+6p\mu}\pi_0 + \frac{2p\mu}{\lambda+6p\mu}\pi_1 + \frac{p\mu}{\lambda+6p\mu}\pi_2 + \frac{2p\mu}{\lambda+6p\mu}\pi^{(1)}_2 \\ \pi_2 = \frac{p\mu}{\lambda+6p\mu}\pi_1 + \frac{p\mu}{\lambda+6p\mu}\pi_3 + \frac{2p\mu}{\lambda+6p\mu}\pi^{(1)}_2 + \frac{p\mu}{\lambda+6p\mu}\pi^{(1)}_3 \\ \pi_{\text{thr}-1} = \frac{p\mu}{\lambda+6p\mu}\pi_{\text{thr}-2} + \frac{p\mu}{\lambda+6p\mu}\pi^{(1)}_{\text{thr}-1} \\ (\forall 3 \le i \le \text{thr}-2) \\ \pi_i = \frac{p\mu}{\lambda+6p\mu}\pi_{i-1} + \frac{p\mu}{\lambda+6p\mu}\pi_{i+1} + \frac{2p\mu}{\lambda+6p\mu}\pi^{(1)}_{i-1} + \frac{p\mu}{\lambda+6p\mu}\pi^{(1)}_{i+1} \end{cases} \tag{4}$$

where $\lceil x \rceil$ is the smallest positive integer greater than or equal to x, and

$$\begin{cases} \pi^{(1)}_2 = \frac{2p\mu}{\lambda+6p\mu}\pi_1 + \frac{2p\mu}{\lambda+6p\mu}\pi_2 + \frac{p\mu}{\lambda+6p\mu}\pi^{(1)}_3 \\ \pi^{(1)}_3 = \frac{2p\mu}{\lambda+6p\mu}\pi_2 + \frac{2p\mu}{\lambda+6p\mu}\pi_3 + \frac{2p\mu}{\lambda+6p\mu}\pi^{(1)}_2 + \frac{p\mu}{\lambda+6p\mu}\pi^{(1)}_3 \\ \quad + \frac{p\mu}{\lambda+6p\mu}\pi^{(1)}_4 + \frac{2p\mu}{\lambda+6p\mu}\pi^{(2)}_4 \\ \pi^{(1)}_4 = \frac{2p\mu}{\lambda+6p\mu}\pi_3 + \frac{2p\mu}{\lambda+6p\mu}\pi_4 + \frac{p\mu}{\lambda+6p\mu}\pi^{(1)}_3 + \frac{p\mu}{\lambda+6p\mu}\pi^{(1)}_5 \\ \quad + \frac{2p\mu}{\lambda+6p\mu}\pi^{(2)}_4 + \frac{p\mu}{\lambda+6p\mu}\pi^{(2)}_5 \\ (\forall 5 < i < \text{thr}-1) \\ \pi^{(1)}_i = \frac{2p\mu}{\lambda+6p\mu}\pi_{i-1} + \frac{2p\mu}{\lambda+6p\mu}\pi_i + \frac{p\mu}{\lambda+6p\mu}\pi^{(1)}_{i-1} + \frac{a(p\mu)}{\lambda+6p\mu}\pi^{(1)}_{i+1} \\ \quad + \frac{p\mu}{\lambda+6p\mu}\pi^{(2)}_i + \frac{a(p\mu)}{\lambda+6p\mu}\pi^{(2)}_{i+1} \end{cases} \tag{5}$$

where

$$a = \begin{cases} 1, & \text{if } 5 \le i \le \text{thr}-2 \\ 0, & \text{if } i = \text{thr}-1 \end{cases}$$

$$\begin{cases} \forall\,(6 < i < (\text{thr}-1) \text{ and } 2 \le j \le \lceil\frac{i-1}{2}\rceil - 1 \\ \pi^{(j)}_i = \frac{p\mu}{\lambda+6p\mu}\pi^{(j-1)}_i + \frac{b_1 p\mu}{\lambda+6p\mu}\pi^{(j+1)}_i + \frac{p\mu}{\lambda+6p\mu}\pi^{(j-1)}_{i-1} \\ \quad + \frac{p\mu}{\lambda+6p\mu}\pi^{(j)}_{i-1} + \frac{b_2 p\mu}{\lambda+6p\mu}\pi^{(j)}_{i+1} + \frac{b_2 p\mu}{\lambda+6p\mu}\pi^{(j+1)}_{i+1} \end{cases} \tag{6}$$

where

$$b_1 = \begin{cases} 1, & \text{if } i \text{ is odd.} \\ 1, & \text{if } i \text{ is even and } 2 \le j \le \lceil\frac{i-1}{2}\rceil - 2 \\ 2, & \text{if } i \text{ is even and } j = \lceil\frac{i-1}{2}\rceil - 1 \end{cases}$$

$$b_2 = \begin{cases} 0, & \text{if } 6 \le i \le \text{thr}-2 \\ 1, & \text{if } i = \text{thr}-1 \end{cases}$$

$$\begin{cases} (\forall 2 \le l \le \lceil\frac{\text{thr}-1}{2}\rceil) \\ \pi^{(l)}_{2l} = \frac{p\mu}{\lambda+6p\mu}\pi^{(l-1)}_{2l} + \frac{p\mu}{\lambda+6p\mu}\pi^{(l-1)}_{2l-1} + \frac{c_1 p\mu}{\lambda+6p\mu}\pi^{(l)}_{2l+1} \end{cases} \tag{7}$$

where

$$c_1 = \begin{cases} 0, & \text{if } l = \frac{\text{thr}-1}{2} \\ 1, & \text{otherwise} \end{cases}$$

$$\begin{cases} \forall 2 \le l \le \frac{\text{thr}-2}{2} \\ \pi^{(l)}_{2l+1} = \frac{p\mu}{\lambda+6p\mu}\pi^{(l-1)}_{2l+1} + \frac{p\mu}{\lambda+6p\mu}\pi^{(l)}_{2l+1} + \frac{p\mu}{\lambda+6p\mu}\pi^{(l-1)}_{2l} \\ \quad + \frac{p\mu}{\lambda+6p\mu}\pi^{(l)}_{2l} + \frac{c_2 p\mu}{\lambda+6p\mu}\pi^{(l)}_{2l+2} + \frac{c_2 p\mu}{\lambda+6p\mu}\pi^{(l+1)}_{2l+2} \end{cases} \tag{8}$$

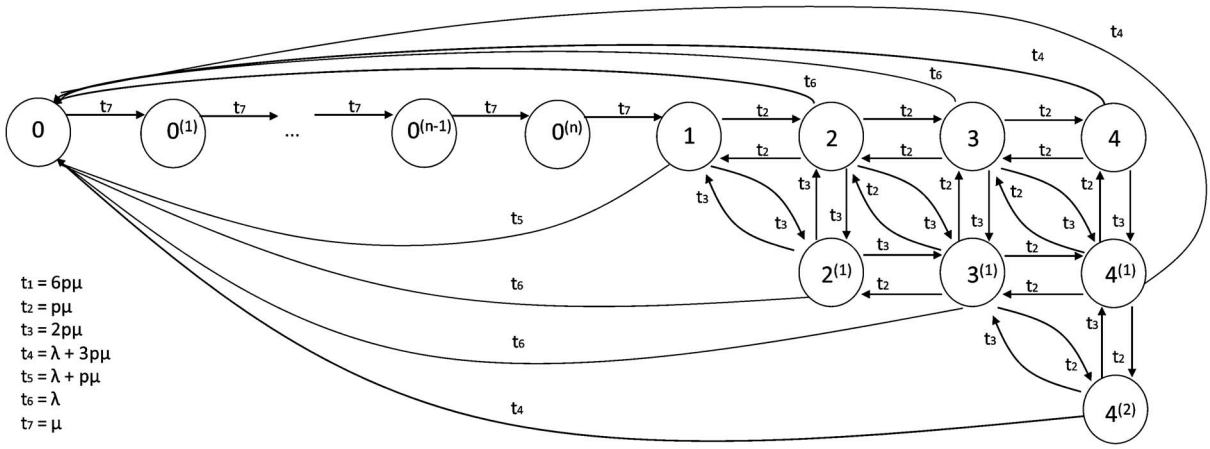

Fig. 10. Modeling Solution 4 using the 2-D mobility model.

where

$$c_2 = \begin{cases} 0, & \text{if } l = \frac{\text{thr}-2}{2} \\ 1, & \text{otherwise} \end{cases}$$

$$\sum_{i=0}^{\text{thr}-1} \pi_i + \sum_{i=2}^{\text{thr}-1} \sum_{m=1}^{\lceil \frac{i-1}{2} \rceil} \pi_i^{(m)} = 1. \tag{9}$$

It shall be noted that (6)–(8) result from the aggregation process. Equation (5) represents the states rising from the first level of aggregation, whereas (6)–(8) relate to another level of aggregation. (For further details on the aggregation process and the used algorithm, see [24].)

For the modeling of Solution 4, we consider the same reasoning as for the precedent solutions. In this solution, the state space increases by including the sub chain $\{0, 0^{(1)}, 0^{(2)}, \ldots, 0^{(n-1)}, 0^{(n)}\}$, which represents n consecutive times that the optimal P-GW is maintained after each handover. Fig. 10 shows the CMTC of Solution 4. Based on this figure, we can derive the steady-state probability of the aggregated states C_i and C_i^m, respectively. The only equations that change or appear from the model of Solutions 2 and 3 are

$$\begin{cases} \pi_0 = \pi_{0^{(1)}} \\ \pi_{0^{(2)}} = \pi_{0^{(1)}} \\ \vdots \\ \pi_{0^{(n-1)}} = \pi_{0^{(n)}} \\ \pi_1 = \frac{6p\mu}{\lambda+6p\mu}\pi_{0^{(n)}} + \frac{2p\mu}{\lambda+6p\mu}\pi_1 + \frac{p\mu}{\lambda+6p\mu}\pi_2 + \frac{p\mu}{\lambda+6p\mu}\pi_{2^{(1)}} \\ \sum_{i=0}^{\text{thr}-1} \pi_i + \sum_{i=0}^{\text{thr}-1} \pi_0^i + \sum_{i=2}^{\text{thr}-1} \sum_{m=1}^{\lceil \frac{i-1}{2} \rceil} \pi_i^{(m)} = 1. \end{cases} \tag{10}$$

B. Performance Metrics

By resolving the systems presented earlier, we can evaluate the performance of the proposed solutions. Two metrics were considered. The first one is the probability that UE is connected to the optimal P-GW. Obviously, for Solutions 2 and 3, this value represents the probability to be in state 0 (π_0). In the case of Solution 4, this probability is derived as follows: $(\pi_0 + \sum_{i=1}^{n} \pi_0^i)$.

The second metric we consider is the cost (COST) of P-GW relocation procedure (UE disconnection). Each solution incurs a cost, in terms of signaling messages, which is important to study. The cost of one UE disconnect procedure is defined as the amount of signaling traffic exchanged between the network and the UE. It is computed as follows:

$$C_{\text{disconnection}} = N_{\text{Nbr-sign-msg}} * \text{Size}_{\text{msg}} \tag{11}$$

where $N_{\text{Nbr-sign-msg}}$ denotes the number of exchanged messages when a P-GW relocation is triggered, and Size_{msg} denotes the average size of signaling messages.

To compute the average cost (noted COST) of each solution, we differentiate between the case of 1-D model and that of 2-D model. In the case of the 1-D model, the cost is computed as follows for Solutions 2 and 3. Recall that the result of Solution 2 can be obtained by setting $\lambda = 0$ and $\text{thr} = k$

$$\text{COST} = \left(\frac{\lambda}{\lambda + 2p\mu} \sum_{i=1}^{\text{thr}-2} \pi_i + \frac{\lambda}{\lambda + 2p\mu} \pi_{\text{thr}-1} \right) * C_{\text{disconnection}}. \tag{12}$$

For Solution 4, the cost is obtained as follows:

$$\text{COST} = \left(\sum_{i=1}^{n} \pi_{0^i} + \frac{\lambda}{\lambda + 2p\mu} \sum_{i=1}^{\text{thr}-2} \pi_i + \frac{\lambda + p\mu}{\lambda + 2p\mu} \pi_{\text{thr}-1} \right) * C_{\text{disconnection}}. \tag{13}$$

Regarding the 2-D model, we follow the same principle by separating the cost of Solutions 2 and 3 from Solution 4. The following equations represent the cost when using Solutions 2 and 3, and the cost when using Solution 4, respectively:

$$\text{COST} = \left[\frac{\lambda}{\lambda + 6p\mu} \left(\sum_{i=1}^{\text{thr}-2} \pi_i + \sum_{i=2}^{\text{thr}-2} \sum_{j=1}^{\lceil \frac{i-2}{2} \rceil} \pi_i^{(j)} \right) + \frac{\lambda + 2p\mu}{\lambda + 6p\mu} \left(\sum_{j=2}^{\lceil \frac{\text{thr}-2}{2} \rceil} \pi_{\text{thr}-1}^{(j)} \right) \right] * C_{\text{disconnection}} \tag{14}$$

$$\text{COST} = \left[\frac{\sum_{i=1}^{n} \pi_{0^i} + \lambda}{\lambda + 6p\mu} \left(\sum_{i=1}^{\text{thr}-2} \pi_i + \sum_{l=2}^{\text{thr}-2} \sum^{\left\lceil \frac{l-2}{2} \right\rceil} \pi_l^{(j)} \right) + \frac{\lambda + 2p}{\lambda + 6p\mu} \left(\sum_{j=1}^{\left\lceil \frac{\text{thr}-2}{2} \right\rceil} \pi_{\text{thr}-1}^{(j)} \right) \right] * C_{\text{disconnection}}. \quad (15)$$

Equations (12)–(15) represent the average COST for the different solutions. The average COST depends on the cost of one disconnection [see (11)] but also on the state when the P-GW relocation is triggered. The latter is stochastic and depends on the probability to be in each state. For instance, in Solutions 2 and 3, the relocation is done only after k distance from the optimal P-GW. This explains why in (12), by replacing λ by zero and thr by k, the average COST will depend only on state $k-1$.

V. Performance Evaluation

Following the description of our proposed schemes, here, we provide a performance comparison of the discussed solutions based on an analytical study using Markov models and through simulations performed via MATLAB. We initially provide an analytical comparison using Markov-based models of the P/S-GW relocation solutions for highly mobile UE using a broad range of values for the update distance threshold, session-to-mobility ratio, session arrival rate, and the number of consecutive relocations. Such a study is then complemented by a set of simulations that evaluate the user context and history-based solution (i.e., Solution 5), where using a Markov-based model is not feasible as this solution is based on user history. Indeed user behavior is easier modeled via simulation, which captures the user habits in a number of prior simulation runs that feed the system with user-context/mobility history data. Simulations are also conducted to evaluate the proposed scheme regarding S-GW relocation avoidance during handover for UE in active mode since the network layout in combination with the user behavior are modeled in a more realistic way.

A. Numerical Results

Here, we present numerical results obtained by resolving the Markov models, which compare the proposed solutions in terms of the probability of UE to be connected to the optimal P-GW and the cost associated with each solution. Based on the 3GPP specifications [4], the number of messages exchanged $(N_{\text{Nbr-sign-msg}})$ between the network and mobile UE is fixed to 12. We assume that each message has an average size $(\text{Size}_{\text{msg}})$ of 10 B. Further, the results are obtained with the following settings: $k = 15$ (Solution 2) and thr $= 15$ (Solutions 3 and 4).

Figs. 11 and 12 show the impact of the threshold distance k on Solution 2 in terms of the probability of UE being connected to the optimal P-GW and the cost associated with P-GW relocation, respectively. The results are for both mobility models.

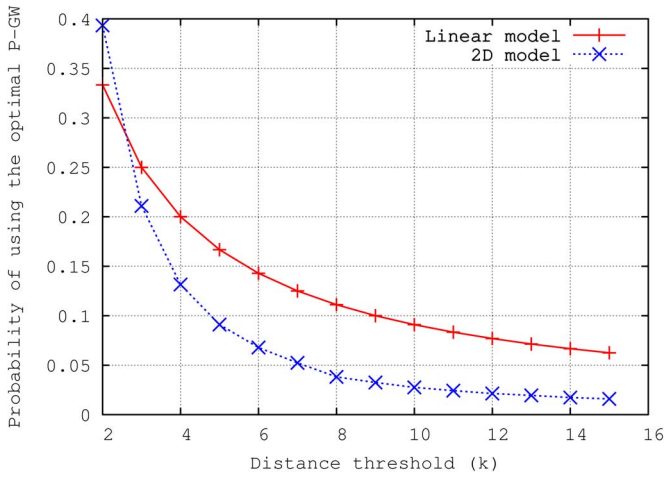

Fig. 11. Impact of distance threshold k on the optimal P-GW connection probability in the distance-based solution (Solution 2).

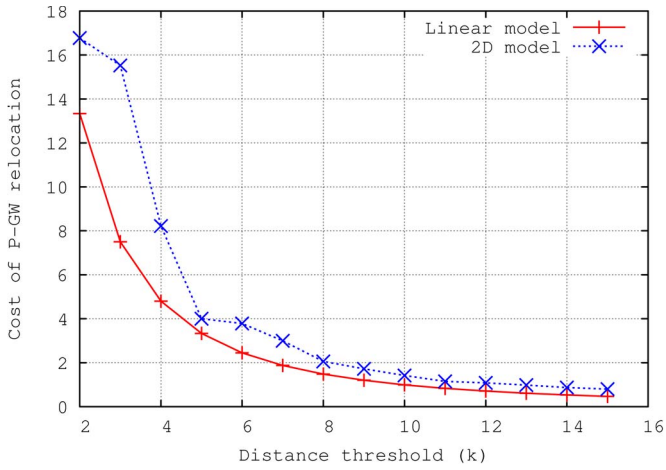

Fig. 12. Impact of distance threshold k on the P-GW relocation cost in the distance-based solution (Solution 2).

We clearly see that both metrics are decreasing functions of k, whereby the lowest probability and cost are obtained for high values of k. This is a consequence of the fact that increasing the distance threshold reduces the cost since disconnections of the UE become less frequent, but at the same, it reduces the probability of the UE being connected to the optimal P-GW as the UE moves far away from the initial/original eNB and the associated P-GW loses its optimality, given the current location of the UE. Accordingly, the value of k defines a tradeoff between reducing the P-GW relocation cost and the user's perceived QoE.

Figs. 13 and 14 show the impact of n on the hybrid solution (i.e., Solution 3) in terms of the optimality of the currently used P-GW and the P-GW relocation cost, respectively. The results are obtained for different values of session-to-mobility ratio (λ/μ). While high values of this ratio mean that the UE is establishing sessions more frequently than performing handoffs, low values represent the case when the UE frequently performs handoffs. Clearly, for high values of n, both the probability of the UE to be connected to an optimal P-GW

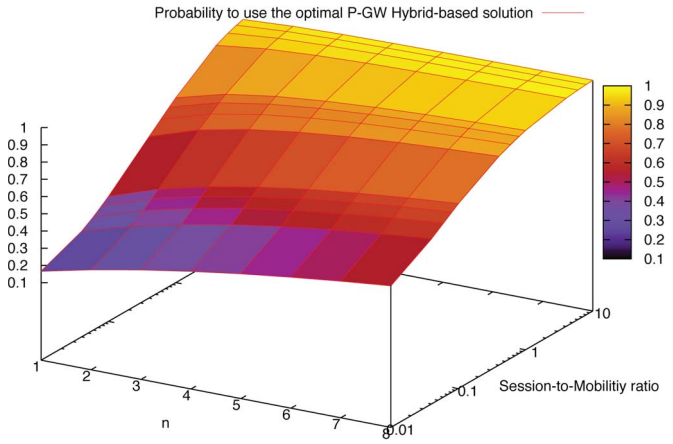

Fig. 13. Impact of n on the optimality of the currently used P-GW in the hybrid solution (Solution 3).

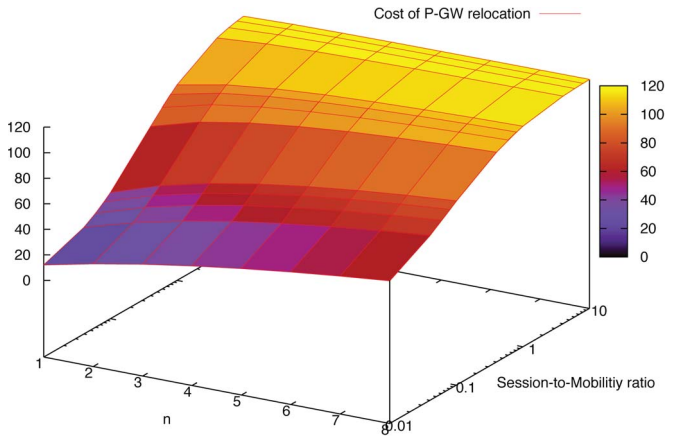

Fig. 14. Impact of n on the P-GW relocation cost in the hybrid solution (Solution 3).

and the cost increase. This performance is expected as, by increasing n, the frequency of P-GW relocations increases, which increases the probability of the UE to be often connected to the optimal P-GW and, hence, the cost. For high values of n, the behavior is identical to Solution 1. Another observation is that increasing the session-to-mobility ratio increases both the probability of the UE to be connected to the optimal P-GW and the cost. This is mainly due to the fact that frequent establishment of IP sessions while residing in the same macrocell increases the overall frequency of relocations to the optimal P-GW.

Figs. 15 and 16 plot the probability of UE to be connected to the optimal P-GW for different session-to-mobility ratios in the case of the 1-D mobility model and the 2-D mobility model, respectively. The two figures compare the results of Solutions 1 (optimal), 2 (distance based), 3 (session based), and Solution 4 (hybrid based). We first remark that Solutions 1 and 2 exhibit constant probability, regardless the value of the session-to-mobility ratio. Solution 1 disconnects UE for each handoff, which always ensures a connection to the optimal P-GW and hence achieves the best results, whereas the worst results are obtained in Solution 2, which is affected only by the threshold k (which is set to a fixed value in these figures).

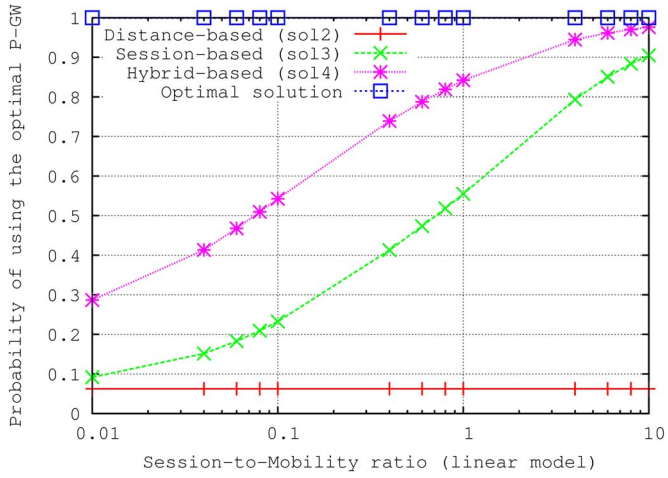

Fig. 15. Probability of UE to be connected to the optimal P-GW for different session-to-mobility ratios. Case of 1-D mobility model.

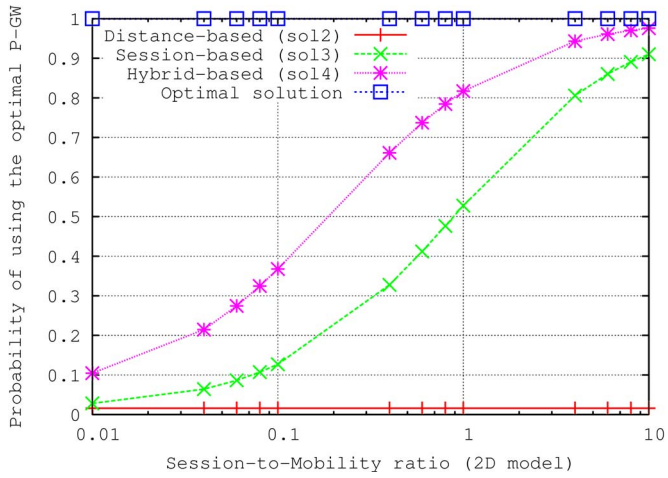

Fig. 16. Probability of UE to be connected to the optimal P-GW for different session-to-mobility ratios. Case of 2-D mobility model.

Moreover, we observe that, compared with Solution 3, Solution 4 achieves the highest probability of UE to be connected to the optimal P-GW as it forces the UE to disconnect for n consecutive handoffs and afterward triggers UE disconnection per IP session establishment. This probability increases along with an increase in session-to-mobility ratio, and reaches the same value as in Solution 1 for high values. In Solution 3, UE connectivity with the optimal P-GW exhibits lower probability compared with Solution 4. However, similar to Solution 4, such probability achieved in the case of Solution 3 increases along with the increase in the session-to-mobility ratio as the frequency of P-GW relocations also increases. The lowest probability is obtained in the case of Solution 2, which triggers P-GW relocation only when the UE is at 15 hops away from the eNB collocated with the currently used P-GW.

Figs. 17 and 18 plot the P-GW relocation cost for different session-to-mobility ratios in the case of the 1-D mobility model and the 2-D mobility model, respectively. The highest cost is incurred in the case of solutions that ensure the highest probabilities of UE to be often connected to the optimal

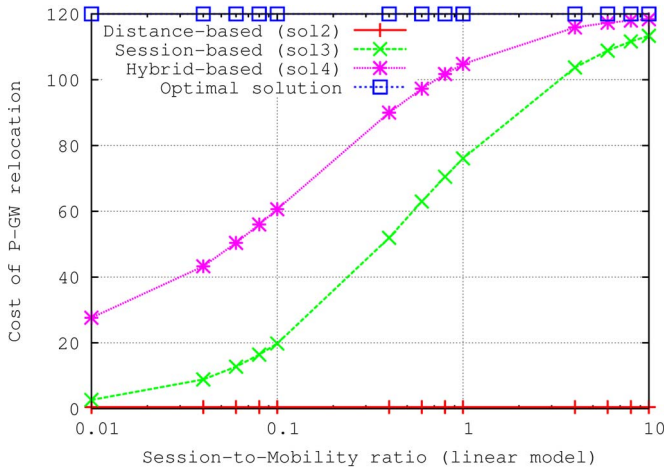

Fig. 17. P-GW relocation cost for different session-to-mobility ratios. Case of the 1-D mobility model.

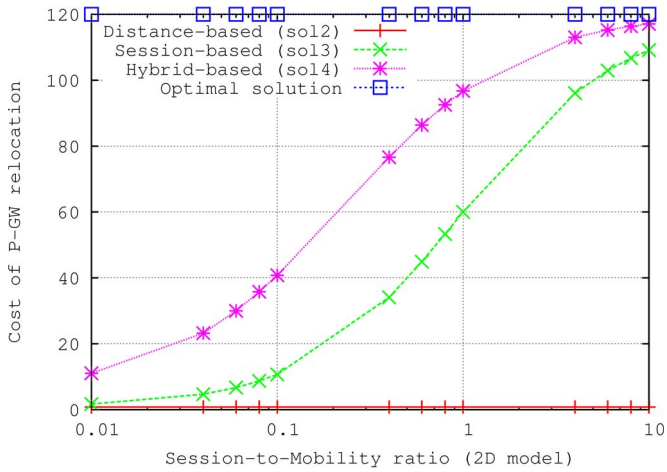

Fig. 18. P-GW relocation cost for different session-to-mobility ratios. Case of the 2-D mobility model.

P-GW. This is intuitively due to the fact that these solutions invoke P-GW relocations more frequently, which incurs higher cost. In the case of Solutions 2 and 3, the cost increases along with an increase in session-to-mobility ratio as these solutions trigger P-GW relocations when a new IP session is initiated. The maximum cost is incurred in Solution 1, whereas the lowest cost is attained in Solution 2.

B. Simulation Results

As previously stated, the simulation study mainly concentrates on evaluating the user-context solution, namely Solution 5, which is based on user behavior and history statistics. In particular, for users in active mode, it is possible to develop their mobility and activity history with fine granularity, at the level of access points. Mobility history can be also obtained for users in idle mode but with a granularity that corresponds to relatively larger areas, such as tracking areas (i.e., set of eNBs), service areas, or MME pool areas in the context of EPS, and routing areas in the context of UMTS. The user-context information is modeled based on monitored user activity from prior simulations. Indeed, we run the simulations multiple times, and for each simulation run, we record the user mobility (as described below) and the locations (i.e., eNB ID) and time instances when users initiate an IP session. In the simulations, an IP session is modeled by its initiation time and its duration (randomly selected from within an interval). We then develop a contextual profile of users, including their mobility patterns and the set of sessions (location, time, and duration) that they initiate. To introduce more realistic scenarios, we considered the difficulty of predicting the user behavior with absolute precision. Hence, errors are introduced deliberately by altering the user-context prediction in terms of mobility and activity, with a certain specified error percentage. To enrich the results, simulations are used to provide a diverse flavor of mobility beyond the random mobility model considered in the Markov-based analysis and to assess the impact of the proposed P-GW solutions under a range of nonuniformly structured network topologies. Regarding the nonrandom mobility model considered in the simulations, we assume users driving around predefined locations, representing vehicle drivers (e.g., taxis or transport vehicle drivers) that perform regular trips among points of attraction. To do so, we assigned points of attractions for each simulated user (UE) and simulated their movement toward these points of attractions at a high probability. We run the simulations multiple times. For each simulation run, we take records of the users' mobility. These records are used to develop (i.e., predict) a mobility pattern that is associated with each simulated user. While the predicted mobility pattern does not always accurately match the actual mobility of a user in every new simulation run, we deliberately introduce further errors in the predicted mobility pattern. This is to simulate relatively realistic scenarios as in real life and to predict, all but impossible, the mobility of users with high accuracy. In addition, we evaluate the proposed S-GW relocation avoidance scheme, which was not considered in the prior analysis and demonstrate its potential in improving the handover performance in the EPS system.

The RAN is formulated as a graph $G(V, E)$ with V eNBs and E indicating adjacency between neighboring cells. To provide a representative result sample considering a dense network arrangement, our study adopts the Erdos–Renyi model [26], wherein the $G(V, p)$ is used to create a random instance for the simulation topology with $V = 80$ number of eNBs and $p = 0.12$ probability of adjacency between two cells. Such a model is adopted to create a series of random and relatively dense RAN topology instances. Such network deployment may represent a network composed by small cells, wherein network decentralization is envisioned. For simplicity, the coverage among neighboring eNBs is assumed to be ideal, assuming no fading and no interference. Effectively, our study and results concentrate solely on the effect of the different GW relocation schemes.

In the following, we assume that 30 UE devices reside in the RAN while being in idle mode, each starting an application session following a Poisson distribution with a mean rate $\lambda = 0.5$ per minute. For the performance evaluation of the different P-GW relocation schemes, we consider a highly decentralized mobile network similar to the one considered in the Markov-based analysis, assuming P-GWs being collocated with eNBs.

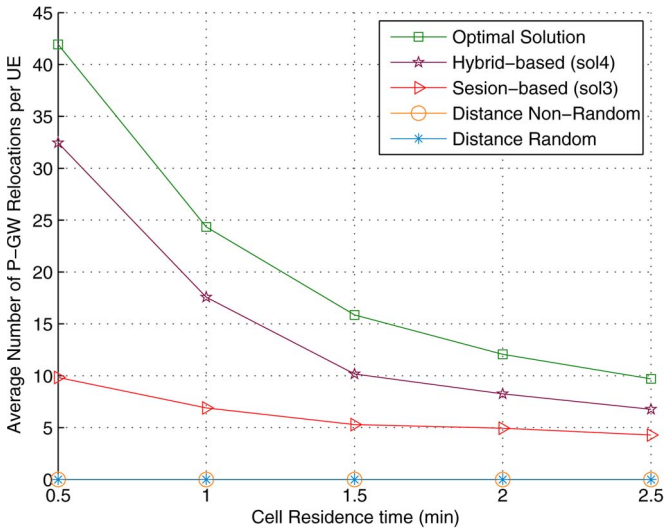

Fig. 19. Performance evaluation of P-GW relocation methods without considering user-context information: signaling overhead due to P-GW relocation.

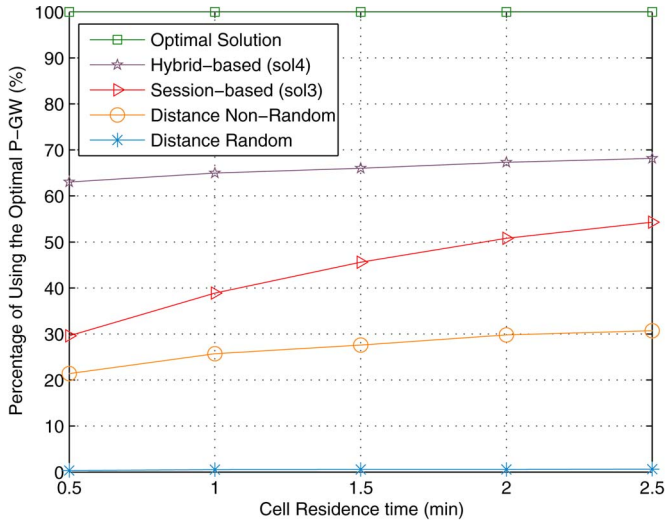

Fig. 20. Performance evaluation of P-GW relocation methods without considering user context information: P-GW optimality.

The simulations were run for a duration of 50 min considering all five P-GW relocation schemes discussed, (i.e., Solutions 1, 2, 3, 4, and 5), in which the update distance $k = 15$ hops for Solution 2 and the consecutive number of optimal updates is thr $= 15$ for Solutions 3 and 4, respectively. As before, the evaluation of the proposed methods is performed considering the tradeoff between the P-GW relocation overhead and P-GW optimality upon initiating an IP session (e.g., probability of UE to be connected to the optimal P-GWs). In particular, the P-GW relocation overhead is measured as the number of P-GW relocations per UE in the entire network, whereas the P-GW optimality is measured as the percentage of UE devices that are associated with the optimal P-GW at the time when they initiate a new IP session. The simulations are performed by altering the mean cell residence time $1/\mu = \{0.5, 1, 1.5, 2.0, 2.5\}$ min to evaluate the performance of the proposed schemes for different mobility speeds. It is important to note that the simulations were run multiple times and the obtained results represent an average of these runs. The confidence interval values of the obtained optimal P-GW percentages was 3%–4% in the case of Solutions 3 and 4, whereas the results obtained in the case of the other approaches exhibited confidence interval values of less than 2%. Considering the overhead measurements, results obtained in the case of all approaches experienced less than 4%–5%. Such values are low and indicate, on one hand, the stability of the simulations and, on the other hand, the validity of the obtained results.

Figs. 19 and 20 show the signaling overhead associated with P-GW relocations and the P-GW optimality for Solutions 1–4, respectively. A general observation is the strong correlation between the P-GW relocation overhead and the P-GW optimality, matching the Markov-based analysis. A high number of P-GW updates produces a high degree of P-GW optimality. Solution 1 ensures constant P-GW optimality irrespective of speed, at the cost of significantly high signaling overhead, whereas the nonoptimal approaches, Solutions 3 and 4, are 30 to 70% less accurate than Solution 2, introducing the minimum overhead at the cost of almost no P-GW optimality. Solution 3 disconnects UE devices and reconnects them back only upon the attempt to establish a new IP session. UE devices need to be then first allocated a P-GW before the actual delivery of data. This intuitively increases the connection establishment delay. All other solutions avoid increasing the connection setup delay since UE devices are always associated with a corresponding P-GW, although they introduce suboptimal routing, which results in increased end-to-end delays. For higher speeds or shorter cell residence times, the difference in the signaling overhead between the optimal approach (Solution 1) and the remaining approaches increases, whereas for lower speeds, all nonoptimal approaches incur a similar amount of signaling overhead, with the exception of Solution 2, which is always zero.

The performance of the distance-based approach, i.e., Solution 2, hinges on the UE's mobility pattern and on the traveled distance threshold after which P-GW relocation occurs. Hence, this solution is beneficial for relatively low number of UE devices, i.e., users that do not communicate frequently, whereas for other users, its performance is proportional to the P-GW update frequency. Effectively, Fig. 20 demonstrates that the distance-based approach (Solution 2) with a random mobility model incurs a lower percentage of using the optimal P-GW compared with the equivalent one, whereby movement is centered around particular locations within the vicinity of the optimal P-GW, despite the fact that both approaches have the same distance limit. In terms of P-GW overhead, both approaches exhibit similar results, as shown in Fig. 19. The hybrid and session-based approaches, i.e., Solutions 3 and 4, incur similar signaling overhead for lower speeds, i.e., when the cell residence time is high, while for higher speeds Solution 4 produce about three time more overhead compared with Solution 3. In terms of P-GW optimality, the hybrid approach outperforms the session-based one. The magnitude of this better performance decreases as the cell residence time increases.

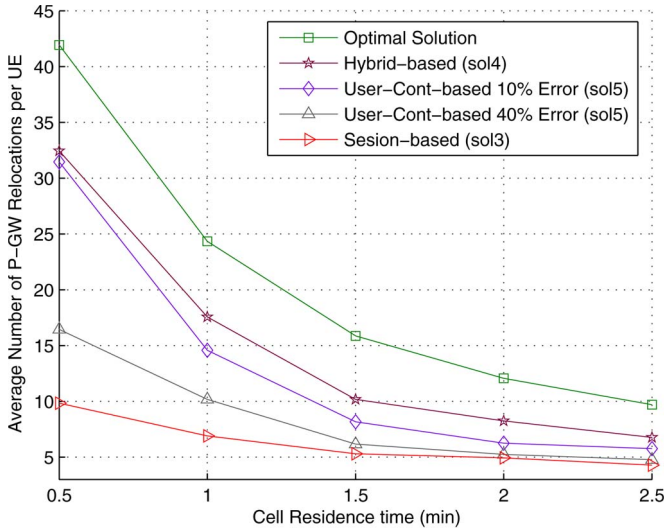

Fig. 21. Enhancing the performance of P-GW relocation considering user context information: signaling overhead due to P-GW relocation.

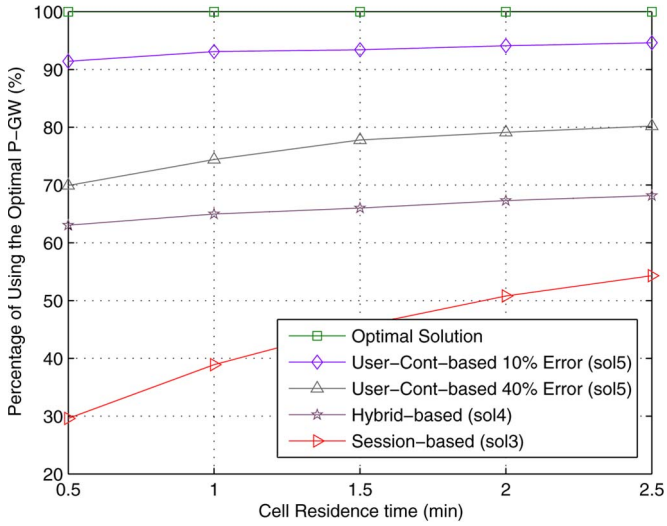

Fig. 22. Enhancing the performance of P-GW relocation considering user context information: P-GW optimality.

The user-context based approach, i.e., Solution 5, provides high P-GW optimality, close to the optimal solution, while resulting in 25%–50% less signaling overhead, i.e., when the information regarding the behavior of particular UE is 90% accurate (see Figs. 21 and 22). However, in practice, it is difficult to predict, with so small error, the behavior of users in terms of IP session establishment. For this reason, we deliberately inserted a higher level of errors in the prediction of the user behavior. Figs. 21 and 22 plot the signaling overhead and the P-GW optimality in the case user behavior is predicted with 40% error. As comparison terms, we use the optimal solution, i.e., the session-based and the hybrid ones. When there are errors in the user behavior prediction, the P-GW optimality degrades when the user-context-based approach is used. This degradation increases along with errors in the prediction. However, the signaling overhead remains comparable to the other schemes. These results demonstrate that with more accurate knowledge on user behavior, an operator can largely improve the usage of its resources.

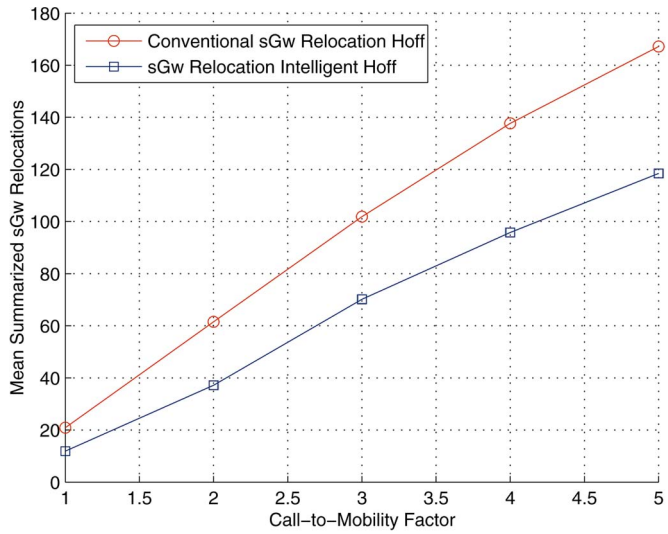

Fig. 23. S-GW relocation frequency for different call-to-mobility factor.

For studying the performance of the S-GW relocation-avoidance-based handover scheme, we assume that neighboring cells overlap creating regions where all participant eNBs exhibit an equal quality of radio characteristics, i.e., UE has similar signal strength. Such conditions introduce an equal handover selection potential for UE residing by the edge of a cell. MME may select, for UE that needs to perform an imminent handover, a random neighbor eNB under the conventional handover scheme (i.e., UE that measures an equal signal strengths); whereas, using the proposed handover scheme, the source eNB commands the associated UE to hand over toward a neighbor eNB within the same service area, avoiding therefore S-GW relocation. A pool of S-GWs is associated with the RAN, assuming that each incoming UE is assigned to the optimal S-GW following the paradigm described in [8], whereby the selected S-GW always resides in the middle of the service area. Once the UE performs a series of consecutive handovers, which occurs after a movement of more than two eNBs away from the initial eNB, the system enforces a S-GW relocation.

Assuming that incoming users arrive independently following a Poisson distribution with an average arrival rate $\lambda = 40$ users per minute and that the cell residence time is exponentially distributed with a mean $d = 2$ min, we vary the mean session holding time, which is also a random exponential variable, within the range of $1/\mu = \{2, 4, 6, 8, 10\}$ min. The underlying mobility model used is a random-based one, and the simulation is run for 20 min, which is long enough to obtain stable results.

Fig. 23 shows the mean number of total S-GW relocations, experienced in the case of the conventional and proposed schemes, for different values of the call-to-mobility factor, which is defined as λ/μ. The figure clearly indicates the superiority of the proposed handover scheme in reducing the number of S-GW relocations. We also observe that the number of S-GW relocations increases along with an increase in the call-to-mobility factor and the average difference between the two

schemes ranges from 15% to 30%. It should be noted that the results depend on the network topology and the underlying mobility scheme, with the random mobility model indicating a typical performance difference.

VI. Conclusion

To cope with the emerging mobile traffic, many mobile operators are interested in decentralizing their networks. This network decentralization will not be effective unless mobility management techniques are also rethought. In this vein, this paper introduced two schemes, particularly designed for users with high-mobility features, such as smartphone users on board vehicles. In one aspect of the paper, solutions were proposed to avoid, whenever possible, unnecessary S-GW/MME relocation during a handoff operation of UE in active mode. For UE in idle mode and traveling for a long distance and/or at a high speed, a number of solutions were proposed to notify these UE devices of the availability of optimal data anchor GWs while minimizing unnecessary PDN disconnections. The proposed optimizations all aim to avoid unnecessary signaling for the PDN reconnection to an optimal data anchor GW and unnecessary application layer signaling for registration/subscription-based applications. The proposed solutions were analytically modeled using a Markov model and considering both the 1-D and 2-D mobility models. The solutions were proven to form a CTMC. Based on these models, the solutions were analyzed, and encouraging results were obtained. From the conducted simulations and the obtained results, it has become clear that approaches that allow the network to simply indicate to UE the availability of an optimal PDN connection, without enforcing PDN disconnection, exhibit a better tradeoff in terms of P-GW optimality and signaling overhead. In some of these approaches, disconnection can still be enforced whenever needed (e.g., at service area boundaries).

Finally, by proposing different possible solutions and evaluating them, we were able to demonstrate the advantages and drawbacks of each, which in turn depend on the underlying scenario; the mobility pattern of the user, i.e., being on board a local or high-speed train; and behavior of the user, i.e., UE being used for web browsing, to watch YouTube videos, to chat, etc. Which solution to use depends, to a large extent, on the operators' policies in how to handle these different possible scenarios.

References

[1] "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2012–2017," White Paper, Feb. 2013. [Online]. Available: http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-520862.html

[2] *Mobile Traffic Growth + Cost Pressures = New Solutions?*, Neu Mobile, Wokingham, U.K., Jan. 2010.

[3] 3rd Generation Partnership Project, TS group services and system aspects; Local IP access and selected IP traffic offload (Rel. 10), 3GPP TR 23.829 V10.0.1, Sophia-Antipolis, France, Oct. 2011.

[4] 3rd Generation Partnership Project, General Packet Radio Service (GPRS) enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) access, TS 23.401, V12.1.0, Sophia-Antipolis, France, Jan. 2013.

[5] 3rd Generation Partnership Project, Architecture enhancements for non-3GPP accesses Rel 10, 3GPP TS 23.402 V12.0.1, Sophia-Antipolis, France, Jun. 2013.

[6] T. Taleb, K. Samdanis, and S. Schmid, "DNS-based solution for operator control of selected IP traffic offload," in *Proc. IEEE ICC*, Kyoto, Japan, Jun. 2011, pp. 1–5.

[7] T. Taleb, Y. Hadjadj-Aoul, and S. Schmid, "Geographical location and load based gateway selection for optimal traffic offload in mobile networks," in *Proc. IFIP Netw.*, Valencia, Spain, May 2011, pp. 331–342.

[8] A. Kunz, T. Taleb, and S. Schmid, "On minimizing SGW/MME relocations in LTE," in *Proc. ACM IWCMC*, Caen, France, Jun. 2010, pp. 960–965.

[9] 3rd Generation Partnership Project, Technical specification group radio access Network, Evolved universal Terrestrial Radio Access Network (E-URAN), Self-configuring and self-optimizing network (SON) use cases and solutions, V9.3.1, TR 36.902, Sophia-Antipolis, France, Apr. 2011.

[10] P. Bosch, L. Samuel, S. Mullender, P. Polakos, and G. Rittenhouse, "Flat cellular (UMTS) networks," in *Proc. IEEE WCNC*, Hong Kong, Mar. 2007, pp. 3861–3866.

[11] Z. Yan, L. Lei, and M. Chen, "WIISE—A completely flat and distributed architecture for future wireless communication systems," in *Proc. WWRF 21*, Stockholm, Sweden, Oct. 13–15, 2008, pp. 1–5.

[12] M. Liebsch, S. Schmid, and J. Awano, "Reducing backhaul costs for mobile content delivery—An analytical study," in *Proc. IEEE ICC*, Ottawa, ON, USA, Jun. 2012, pp. 2895–2900.

[13] C. B. Sankaran, "Data offloading techniques in 3GPP Rel-10 networks: A tutorial," *IEEE Commun. Mag.*, vol. 50, no. 6, pp. 46–53, Jun. 2012.

[14] K. Samdanis, T. Taleb, and S. Schmid, "Traffic offload enhancements for eUTRAN," *IEEE Commun. Surveys Tuts.*, vol. 14, no. 3, pp. 894–896, Sep. 2012.

[15] H. Chan, "Problem statement for distributed and dynamic mobility management," IETF Internet Draft, Jul. 2011.

[16] R. Kuntz, D. Sudhakar, R. Wakikawa, and L. Zhang, "A summary of distributed mobility management," IETF Internet Draft, Feb. 2011.

[17] P. Bertin, S. Bonjour, and J.-M. Bonnin, "Distributed or centralized mobility?" in *Proc. IEEE GLOBECOM*, Honolulu, HI, USA, Dec. 2009, pp. 1–6.

[18] C. Xue, J. Luo, R. Halfmann, E. Schulz, and C. Hartmann, "Inter GW load balancing for next generation mobile networks with flat architecture," in *Proc. IEEE VTC-Spring*, Barcelona, Spain, May 2009, pp. 1–5.

[19] M. Fischer, F.-U. Andersen, A. Kopsel, G. Schafer, and M. Schlager, "A distributed IP mobility approach for 3G SAE," in *Proc. IEEE 19th PIMRC*, Cannes, France, Sep. 2008, pp. 1–6.

[20] K. Kyamakya and K. Jobmann, "Location management in cellular networks: Classification of the most important paradigms, realistic simulation framework, and relative performance analysis," *IEEE Trans. Veh. Technol.*, vol. 54, no. 2, pp. 687–708, Mar. 2005.

[21] M. Toril, "Automatic re-planning of tracking areas," in *Proc. SOCRATES Final Workshop Self-Org. Mobile Netw.*, Karlsruhe, Germany, Feb. 22, 2011, pp. 1–32.

[22] A. Nadembega, A. Hafid, and T. Taleb, "A path prediction model to support mobile multimedia streaming," in *Proc. IEEE ICC*, Ottawa, ON, Canada, Jun. 2012, pp. 2001–2005.

[23] S. Pack, T. Kwon, and Y. Choi, "A performance comparison of mobility anchor point selection schemes in hierarchical mobile IPv6 networks," *J. Comput. Netw.*, vol. 51, no. 6, pp. 1630–1642, Apr. 2007.

[24] R. Langar, N. Bouabdallah, and R. Boutaba, "A comprehensive analysis of mobility management in MPLS-based wireless access networks," *IEEE/ACM Trans. Netw.*, vol. 16, no. 4, pp. 918–931, Oct. 2008.

[25] K.-H. Chiang and N. Shenoy, "A 2-D random-walk mobility model for location-management studies in wireless networks," *IEEE Trans. Veh. Technol.*, vol. 53, no. 2, pp. 413–424, Mar. 2004.

[26] B. Bollobas, *Random Graphs*., 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2001.

[27] T. Taleb, Y. Hadjadj-Aoul, and K. Samdanis, "Efficient solutions for data traffic management in 3GPP networks," *IEEE Syst. J.*, doi: 10.1109/JSYST.2013.228397.

[28] T. Taleb and A. Ksentini, "Follow me cloud: Interworking federated clouds and distributed mobile networks," *IEEE Netw.*, vol. 27, no. 5, pp. 12–19, Sep./Oct. 2013.

[29] T. Taleb and A. Ksentini, "On efficient data anchor point selection in distributed mobile networks," in *Proc. IEEE ICC*, Budapest, Hungary, Jun. 2013, pp. 6289–6293.

[30] T. Taleb, K. Samdanis, and F. Filali, "Towards supporting highly mobile nodes in decentralized mobile operator networks," in *Proc. IEEE ICC*, Ottawa, ON, Canada, Jul. 2012, pp. 5398–5402.

[31] C. Stefanovic, D. Vukobratovic, F. Chiti, L. Niccolai, V. Crnojevic, and R. Fantacci, "Urban infrastructure-to-vehicle traffic data dissemination using UEP rateless codes," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 1, pp. 94–102, Jan. 2011.

Tarik Taleb (SM'10) received the B.E. degree in information engineering (with distinction) and the M.Sc. and Ph.D. degrees in information science from Tohoku University, Sendai, Japan, in 2001, 2003, and 2005, respectively.

From October 2005 to March 2006, he was a Research Fellow with the Intelligent Cosmos Research Institute, Sendai. Then, until March 2009, he was an Assistant Professor with the Graduate School of Information Sciences, Tohoku University, in a laboratory fully funded by KDDI: the second largest network operator in Japan. He is currently working as a Senior Researcher and Third-Generation Partnership Project (3GPP) Standards Expert at NEC Europe Ltd., Heidelberg, Germany. He is leading the NEC Europe Labs Team working on research and development projects on carrier cloud platforms. He has been also directly engaged in the development and standardization of the Evolved Packet System as a member of 3GPPs System Architecture working group. His research interests include architectural enhancements to mobile core networks (particularly 3GPPs), mobile cloud networking, mobile multimedia streaming, congestion control protocols, handoff and mobility management, intervehicular communications, and social media networking.

Dr. Taleb is a Distinguished Lecturer of the IEEE Communications Society (ComSoC). He also served as Secretary and then as Vice Chair of the Satellite and Space Communications Technical Committee of IEEE ComSoc from 2006 to 2010. He currently serves as the Vice Chair of the Wireless Communications Technical Committee of IEEE ComSoC. He is also a Board Member of the IEEE ComSoc Standardization Program Development Board. As an attempt to bridge the gap between academia and industry, he has founded and has been the General Chair of the IEEE Workshop on Telecommunications Standards: from Research to Standards, which is a successful event that received the Best Workshop Award from the IEEE ComSoc. He has been on the editorial board of the IEEE WIRELESS COMMUNICATIONS MAGAZINE, the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, *IEEE Communications Surveys & Tutorials*, and a number of Wiley journals. He has been on the technical program committee of different IEEE conferences, including the IEEE Global Communications Conference, the IEEE International Conference on Communications, and the IEEE Wireless Communications and Networking Conference, and chaired some of their symposia. He received the Young Researcher's Encouragement Award from the Japan Chapter of the IEEE Vehicular Technology Society in 2003, the Niwa Yasujirou Memorial Award in 2005, the IEEE Computer Society Japan Chapter Young Author Award in 2006, the FUNAI Foundation Science Promotion Award in 2007, the TELECOM System Technology Award from the Telecommunications Advancement Foundation in 2008, and the IEEE ComSoc Asia–Pacific Best Young Researcher Award in 2009. He also received several Best Paper Awards for some of his work.

Konstantinos Samdanis (M'13) received the Ph.D. degree in mobile communications from Kings College London, London, U.K.

He is currently a Senior Researcher and a Backhaul Standardization Specialist with NEC Europe, Heidelberg, Germany. He is the Leader of a research project on Long-Term Evolution network management and of the network virtualization Work Package in the Seventh Framework Programme Initial Training Network CROSSFIRE Marie Curie Action. His standardization activities also include software-defined networking and cloud for broadband multiservice networks.

Dr. Samdanis served as an Editor for the IEEE COMMUNICATIONS MAGAZINE and the *IEEE Communications Society Multimedia Communications Technical Committee Society E-Letters*. He is the Editor of the Energy Efficient Mobile Backhaul work item in the Broadband Forum. He is the Next-Generation Networking Symposium Co-Chair of the 2014 IEEE International Conference on Communications.

Adlen Ksentini (M'11) received the M.Sc. degree in telecommunication and multimedia networking from the University of Versailles, Versailles, France, and the Ph.D. degree in computer science from the University of Cergy-Pontoise, Cergy-Pontoise, France in 2005, with a dissertation on quality-of-service (QoS) provisioning in IEEE 802.11-based networks.

He is currently an Associate Professor with the University of Rennes 1, Rennes, France. He is a member of the INRIA Rennes team Dionysos. He is involved in several national and European projects (FP7 Alicante and FP6 Anemone) on quality-of-service (QoS) and quality-of-experience (QoE) support in future wireless and mobile networks. Recently, he launched a bilateral collaboration with Orange Labs on small cell networks. He is an author or co-author of over 60 papers in technical journals and international conferences. His research interests include future Internet networks, cellular networks, green networks, QoS, QoE, and multimedia transmission.

Dr. Ksentini has served on the Technical Program Committees of major IEEE Communication Society Conferences, including the IEEE International Conference on Communications (ICC), the IEEE Global Communications Conference, the IEEE International Conference on Multimedia and Expo, the IEEE Wireless Communications and Networking Conference, and the IEEE International Conference on Personal Indoor and Mobile Radio Communications. He received the Best Paper Award at the Association for Computing Machinery Symposium on Modeling, Analysis, and Simulation of Wireless and Mobile Systems in 2005 and at the IEEE ICC in 2012.